This article has been accepted for publication by Edinburgh University Press in Corpora. Volume 10, Issue 1, Page 83-94. April 2015.
http://dx.doi.org/10.3366/cor.2015.0067

WordWanderer: A Navigational Approach to Text Visualisation

Marian Dörk[1] and Dawn Knight[2]

Text visualisations provide visual representations of documents or small corpora with the primary aim of supporting language analysis. We are interested in developing a more playful approach to language that can be characterised by the notion of wandering as an open-ended movement. To support such a casual form of engagement with text, we designed the WordWanderer system: a visualisation technique that extends tag clouds into a navigational interface for text. The tool supports the gradual movement between word 'context views', which represent the words that co-occur in the vicinity of the selected word, and word-'comparison views', which arrange words based on their association strengths between two selected words. We report on the encouraging feedback from a ten-day deployment of the interface and present promising directions for future design and research.

Keywords Corpus linguistics; text analysis; information visualisation; digital humanities; exploratory search.

1. Introduction

Tag clouds are a simple but effective way of representing the distribution of words in a document or corpus. They are widely employed for both casual use and serious analysis (Viégas and Wattenberg, 2008). In particular, the low entry barrier to customisation and sharing through tools such as Wordle expanded their participatory potential (Viégas et al., 2009). They are also the basis for advanced text visualisation techniques that have been developed, for example, to support comparative analysis of corpora (Collins et al., 2009) and to explore the temporal dynamics of tags (Lee et al., 2010). Furthermore, tag clouds have been integrated into analysis environments developed to support, among other tasks, the visual exploration of named entities in literary texts (Vuillemot et al., 2009). The WordSeer system provides interactive access to a range of text processing tools including several visualisations that support scholarly approaches to literary text (Muralidharan and Hearst, 2013). One of the visualisations included in WordSeer is word tree, a technique that transforms text into a hierarchical representation based on a selected word or phrase (Wattenberg and Viégas, 2008). Similarly, the phrase nets technique transforms an unstructured text into a graph of words connected with each other on the basis of a text pattern (van Ham et al., 2009). Visualisation has also found use in corpus linguistics, the study of real life, naturally occurring language data. There is a growing selection of software environments for corpus analysis that are feature-rich and sophisticated tools, aimed mainly at people with expertise in linguistics (see Scott, 1999 and Rayson, 2003 for examples). The purpose of this research is to explore the potential of interactive text visualisations for language analysis by non-experts.

[1] Institute for Urban Futures, Potsdam University of Applied Sciences, Pappelallee 8–9, 14469, Potsdam, Germany.
[2] School of English, Communication and Philosophy, Cardiff University, Cardiff, CF10 3EU, UK.
Correspondence to: Marian Dörk, *e-mail*: doerk@fh-potsdam.de

2. Towards a navigational approach to text

We are interested in supporting a playful approach to language that involves casual investigations and does not require prior knowledge of linguistics or visualisation. The success of tag clouds has demonstrated the great potential that visualisation can have to help people with varying backgrounds to explore language. Encouraged by this development, we wish to go beyond the construction of static visualisations, and support a more dynamic engagement with text that incorporates some of the functionality from corpus based tools. We are inspired by the information flaneur who makes sense of information spaces merely by curiously traversing them (Dörk et al., 2011). The notion of strolling has already been applied to the visualisation of faceted collections (Dörk et al., 2012) – that is to say, a corpus of documents that share a set of attributes such as authors, keywords and citations. With this work we adopt the attitude of a flaneur and aim to formulate a navigational approach to the exploration of unstructured text.

To develop such a navigational approach to text we draw from visualisation techniques and on methods that are used in language description and analysis, including corpus-based study. Typically, corpus-based research focuses on defining and exploring recurring patterns in language use. If we wish to understand the patterns of use for a particular word we first need to find out how common it is in a language. Given this, the typical 'way-in' to the analysis of corpora is through generating frequency lists, to map out, compare, and contrast how often particular word forms and/or phrases occur across an entire corpus or particular sub-corpora. Beyond frequency counts, explorations of key collocates of search terms are often used as the basis for mapping patterns of word use to gain a better understanding of their roles and functions in discourse.

Our intention is to design a simple yet powerful way of exploring language patterns with potential users of any age and background–from the expert to the intrigued novice. Therefore, while the tool may draw on the knowledge and expertise of language analysts, it is not targeted at this user-group. Existing text analysis tools and visualisations are either results of complex series of data operations, or are created and customised as more or less static representations. We do not wish to frame the interaction with a given text as a primarily analytical process that involves filtering, adjusting parameters, and applying transformations. Instead, we wish to support a navigational mode of analysis that encourages the viewer to move casually through a text.

3. Elements of the design

3.1. Designing for playful text analysis

An iterative design process was undertaken over a period of four months. This process started with general ideas about text analysis leading to initial sketches of visual interfaces for corpus analysis, and resulted in a web-based text visualisation. The general purpose of this visualisation is to encourage a playful approach to corpus analysis by supporting gradual movements through the text. Before discussing the design decisions, consider the example of the fairy tale, Hansel and Gretel. As shown in Figure 1 to 4 the tag cloud reveals the main protagonists Hansel and Gretel (a). Moving the mouse pointer over 'forest' shows that 'children' often co-occurs with that word (b). After the analyst chooses 'forest', its collocates are organized according to their relative proximity in the text (c). Dragging a line between 'children' and 'forest' activates the comparison view (d), which arranges collocates according to their relative strength of association to each of the two selected words. Moving the mouse pointer over the visualisation will trigger gradual highlights indicating varying strengths of co-occurrence, however, the changes in colour and intensity are slightly animated to avoid jarring flicker effects. The instance list is linked with the visualisation; this allows for the highlighting and selection of words from the list as well.

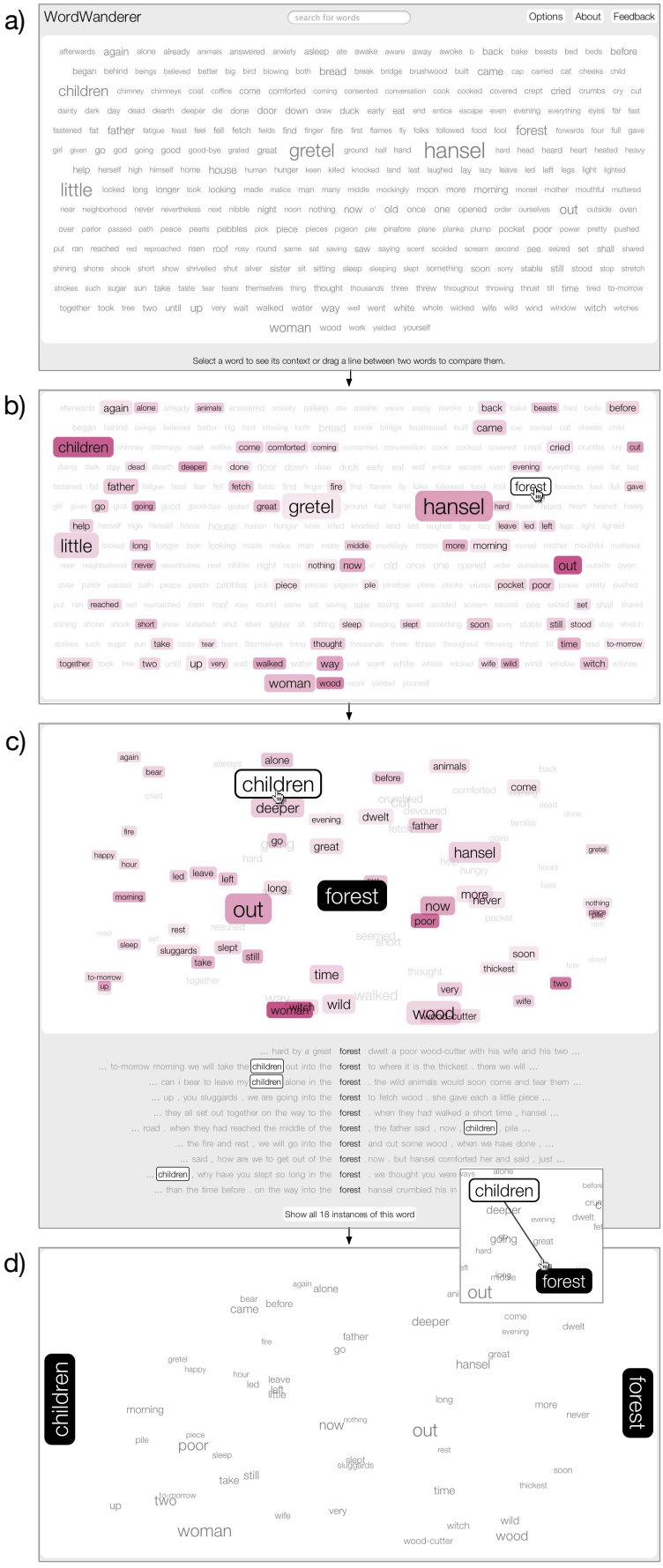

Figure 1: Navigating between different views of a well-known fairy tale in WordWanderer.

3.2. Visualising word associations

The main part of the WordWanderer interface comprises the visualisation that can have one of three states: a cloud (no word selected), context (one word selected), and comparison view (two words selected). Given that font size has been identified as one of the particularly useful visual variables for the design of tag clouds (Bateman et al., 2008), we use it to encode overall frequency in the cloud view and association strength in the context and comparison views. The vertical positioning of the words is consistently alphabetical in all views, while the horizontal positioning varies depending on the mode.

The *cloud view* uses a layout that resembles most tag clouds: an alphabetical ordering from left to right, top to bottom as the text flows (see Figure 1, top left). Given that tag clouds are the most widely used form of visualisation on the web, we chose it as the initial view to provide an accessible and interpretable entrance point for the other more advanced views.

The *context view* is a concordance visualisation that places the selected word (node) into the centre of the view and arranges the collocates (associated words) horizontally according to their aggregated relative position. For example, the word 'children' appears more often before 'forest' in the fairy tale Hansel and Gretel and is, therefore, arranged more towards the left side (see Figure 1, bottom right). The font size of a collocate represents the association strength with the node, which is based on the number of co-appearances and their respective proximities in the text. The association strength for a node *n* and collocate *c* is calculated as the weighted sum of co-occurrences *F(n,c,d)* each for a given distance *d* within a span of seven collocates both before and after the node:

$$s(n,c) = \sum_{d=1}^{7} \left(\frac{F(n,c,d)}{d} \right)$$

With this custom collocation measure, the association strength is higher the more often collocates occur and the closer these occurrences are to the node. In addition, below the context view a concordance list shows the instances of the selected term and its co-text.

The *comparison view* is based on the selection of two words, displaying those words that are collocates with both selected words. While the font size is based on an average of the two association values, the difference in association between the two selected words is expressed by the horizontal positioning of the collocates. Words that are more associated with the left word are positioned more towards the left side and vice versa. For example, the collocate 'woman' appears more towards the left side and this indicates that it is more strongly linked with 'children' than with 'forest' (see Figure 1, c).

There is not a fixed cut-off for a given word frequency or association strength, but rather a maximum number of words that can be displayed based on the layout: The cloud view shows the 300 most frequent words, while the context and comparison views show the 100 words with the highest association strengths. In effect, the minimum frequency or association strength is dynamic and depends on the selected text.

3.3. Wandering between words

We designed the interactivity towards simplicity to encourage a navigational engagement with the text. The basic interaction methods are highlighting, selection, and search.

- *Highlighting.* In any view, one can move the mouse pointer over a word to see the varying association strengths of the remaining words indicated as pink highlights (see Figure 1, b and c). The highlights vary in contrast such that the stronger the association a given word has with the hovered-over word the darker the tone of pink

used for the background of the respective word. When a word does not co-occur with the hovered-overed word, it will be shown in a light grey.

- *Selection*. To select a single word one simply clicks on it. The word will move into the middle of the visualisation and the respective collocates will transition according to their overall positions relative to the selected word in the text. To select two words for the comparison view, one drags a line between the two words. When the mouse button is released, the two words transition to the sides of the view. The collocates will move to their respective positions according to their comparative collocation values. One can return to the cloud view either by removing a word as a selection by clicking on it again or by pressing the escape key.
- *Search*. Besides selecting words by clicking on them, it is also possible to undertake a full-text search to select a word quickly or to find less common words not appearing in the visualisation. After entering a few letters a list of word suggestions appears beneath the search box. Highlighting the results, either using the keyboard or the mouse, accentuates the respective collocates and the selected word in the visualisation (if present). By selecting an item from the result list the word becomes the new basis for the visualisation.

By default, the visualisation excludes common words such as the, is and to, as well as punctuation marks, symbols and numbers. However, an options panel allows the viewer to select the word types that they are interested in. Furthermore, they can enter words that they wish to exclude from the visualisation; this is especially useful if such words are likely to be particularly frequent in a text and thus have the potential to dominate the rest of the visualisation.

4. Deployment

4.1. Launching the WordWanderer prototype

To explore the potential of the WordWanderer interface for text analysis, a prototype of the tool was deployed as a web application. The interface allowed the visitor to enter their own text or choose from five sample texts, one of which was Hansel and Gretel. At present, WordWanderer only uses individual texts of up to about 50,000 words (depending on the visitor's machine's performance), but it can be extended to much larger corpora when coupled with a backend and a database. The site was launched during an annual applied linguistics conference during which a link had been tweeted with the conference hashtag, flyers were distributed to delegates (including publishers, students and academics), and five small focus groups (involving nine participants in total) were called on to provide feedback. Even though the tool is eventually targeted at people with a limited background in linguistics, we first reached out to the linguist community to ensure that linguist methods are appropriately implemented and exposed. Interest in the tool was also generated through social media networks, academic mailing lists and personal contacts. Users of the interface were prompted to provide a few comments on its potential utility, applications and improvements.

4.2. Reception

Within ten days of launching the site, over 1,800 'wandering' sessions were recorded. In this period, the link to the software was also tweeted through the site fifty-three times, and liked on Facebook fifty-three times, too. The software has also been posted onto a corpus linguistics Facebook page, where it has received a further twenty-three 'likes'. In total, seventy-six different feedback comments were received across these different communicative modes, and these came from a range of individuals, including programmers, teachers, academics, students and publishers.

The majority of this feedback commented on the tool in a positive way with respondents noting that the tool was 'interesting' (fourteen times), 'great' (twelve), 'fun' (eight), 'good' (seven), 'brilliant' (four), 'easy to use' / 'user friendly' (five), 'cool' (three) and 'fascinating' (three). These were the most commonly used positive adjectives in the feedback obtained. Less positive comments described the interface as 'blurry' (two), too 'small' or 'tiny' (two), 'unclear' (two), with overlapping words (two), making it 'hard' / 'difficult' (six) to understand and interpret.

These are aspects that will be revised in the next iteration of the tool. For example, the visualisation should be made responsive to take into account the size of browser and to utilise all available screen space. Furthermore, overlaps between words should be minimised to ease the readability of the displayed words and, thus, the interpretability of the arrangement. Several visitors commended the visual appearance, with comments such as 'the interface is very slick and the visual design is very nice', and 'love the viz; very elegant'. The comparison view, created by dragging a line between two words, provided the source of extensive commendation from visitors with regard to its simplicity and interactivity, while the context view was noted as being 'very helpful and inspiring'.

4.3. Applications

Regarding possible usage and applications, the majority of those providing feedback commented that it has a potential use in teaching and learning at all levels–from young learning, to secondary (A-level) students, English as a second language and graduate students. A focus-group participant specifically questioned 'Is it a game? It looks like a game' and noted that 'kids work with visuals, [so] why can't adults?' Another respondent recommended extending the format of the tool to allow for students to perform group work on touch screens and interactive whiteboards in classroom contexts.

Furthermore, a focus-group participant suggested that the software has potential value for 'language awareness and literacy not just language learning', especially in light of newly introduced grammar, punctuation and spelling tests[3] and closer attention to grammar tuition and literacy within the UK National Curriculum as a whole, is a particularly invaluable potential future application of the software.

Considering closer attention to grammar tuition and literacy, language awareness is a particularly invaluable and resonant potential future application of the software. The accessible structure of the WordWanderer interface lends itself to a focus on form, and, thus, may enable us to tackle this question head-on, potentially providing a utility for developing research skills in users, and for supporting self-organised learning (Thomas and Harri-Augstein, 1985). The effectiveness of this tool in achieving these aims is something we intend to investigate, utilising iterative, participatory processes to design, implement and evaluate the software.

Other suggested potential applications include its use as a tool for language description in general, for comparing 'scientific texts in the same domain', analysing political speeches, 'visualizing a text from different perspectives to highlight certain aspects', and for exploring narratives in texts (from a literary–linguistic perspective). We will investigate the potential of adapting the tool for each of these uses, and others, over the next twelve months of our research.

[3] For more information on these governmental reforms visit:
http://media.education.gov.uk/assets/files/pdf/2/sta136001_2013%20ks2%20ara.pdf

4.4. Future iterations and improvements

In addition to the potential of the WordWanderer interface, the feedback we received also highlights a wide range of useful areas for improvement and extension of the tool. The suggestions that were mentioned most often are as follows:

- Include a way to save and share wanders using permalinks.
- Allow users to compare multiple texts.
- Add clearer mapping on the context view to show the relationship between words at the sentence level.
- Include a tool for calling up all instances of a word family (operating at the lemma rather than word level).
- Support the browser's back button.

These will act as the starting points for future developments of the software. In terms of functionality, some users found it difficult to determine the specific purpose of the tool, what it is doing and why, and who it might be aimed at, with some observing that they experienced some difficulty in gaining any real insight from exploring the texts due to the unstructured approach to the website (fourteen respondents provided such comments). Approaching texts from the word, rather than sentence, level was the source of much of this confusion, and this is something that can potentially be addressed in the next iteration of the tool; and more clarity regarding the purpose and functionality of the tool needs to be provided.

5. Discussion and conclusion

With the WordWanderer prototype tool we have aimed to make the following contributions:

- An approach to text analysis that complements the use of overviews with the navigation between partial views.
- A novel visualisation interface that is designed to support a playful engagement with a text corpus.
- Initial observations and feedback from a two-week deployment.

The range of people we targeted with the prototype is perhaps still a little narrow in terms of expertise, and these people are likely to approach and assess the site in a different way from the average layperson. However, the interest generated and the positive feedback received provide us with invaluable signposts for future developments of design and infrastructure, as well as a clearer understanding of the potential applications of this tool.

In particular, language learning has been suggested several times as an area where the approach taken by the WordWander could prove to be beneficial. In fact, language learning has already been identified as a promising area for corpus-based approaches that contribute to an improved understanding of language.

'Teaching to exploit' real-life language data (as with corpus-based approaches to pedagogy – for example, see Carter et al., 2000; Johns, 1991; McCarthy et al., 2005; McEnery et al., 2006; and Thornbury, 2004) in language learning contexts is a process which perhaps provides the most autonomy for learners of a language, as this is a process that is 'necessarily inductive' whereby 'learners inspect the evidence and look for patterns in the data from which they can form generalisations' (Thompson, 2006: 10). The question of how we can best facilitate this process, to encourage and support 'student's habits of observation, noticing, or conscious exploration of grammatical forms and function' (Carter, 1998: 51), using real-life language data is an on-going one. This is potentially something that future iterations of the WordWanderer might start to address.

The WordWanderer is free to access and use and is available online.[4]

[4] See: www.wordwanderer.org

References

Bateman, S., Gutwin, C., and Nacenta, M. (2008). Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the 19th ACM conference on Hypertext and hypermedia, ACM*: 193–202.

Carter, R. (1998). Orders of reality: CANCODE, communication, and culture. *ELT journal 52*(1): 43–56.

Carter, R., Hughes, R., & McCarthy, M. (2000). *Exploring Grammar in Context.* Cambridge: Cambridge University Press.

Collins, C., Viégas, F. B., and Wattenberg, M. (2009). Parallel tag clouds to explore and analyze faceted text corpora. In *VAST 2009: IEEE Symposium On Visual Analytics Science And Technology, IEEE*: 91–98.

Dörk, M., Carpendale, S., and Williamson, C. (2011). The information flaneur: A fresh look at information seeking. In CHI *'11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM* : 1215–1224.

Dörk, M., Riche, N. H., Ramos, G., and Dumais, S. Pivotpaths: Strolling through faceted information spaces. (2012). *TVCG: Transactions on Visualization and Computer Graphics 18*(12): 2710–2719.

Johns, T. (1991). "Should you be persuaded": Two samples of data-driven learning materials, in Johns, T. and King, P. (Eds.) *Classroom Concordancing*. Birmingham: Centre for English Language Studies, University of Birmingham.

Lee, B., Riche, N. H., Karlson, A. K., and Carpendale, S. (2010). Sparkclouds: Visualizing trends in tag clouds. *TVCG: Transactions on Visualization and Computer Graphics 16*(6): 1182 –1189.

McCarthy, M. J., McCarthy, J. and Sandiford, H. (2005) *Touchstone. Student's Book 1*. Cambridge: Cambridge University Press.

McEnery, T. & Xiao, R. and Tono, Y. (2006). *Corpus-based language studies: an advanced resource book.* London: Routledge.

Muralidharan, A., and Hearst, M. A. (2013). *Supporting exploratory text analysis in literature study. Literary and Linguistic Computing 28*(2): 283–295.

Rayson, P. (2003). Matrix: *A Statistical Method and Software Tool for Linguistic Analysis Through Corpus Comparison*. PhD thesis, Lancaster University.

Scott, M. (1999). *Wordsmith tools [computer program]*. Oxford: Oxford University Press.

Thomas, L. F., and Harri-Augstein, E. S. (1985). *Self-organised learning: Foundations of a conversational science for psychology*. Routledge and Kegan Paul: London.

Thompson, P. (2005). Spoken language corpora. In *Developing Linguistic Corpora: A Guide to Good Practice.* Oxford: Oxbow Books, M. Wynne, Ed. Oxbow Books, Oxford, pp. 59–70.

Thornbury, S. (2004) *Natural Grammar*. Oxford: Oxford University Press.

Van Ham, F., Wattenberg, M., and Viégas, F. B. (2009). Mapping text with phrase nets. *TVCG: Transactions on Visualization and Computer Graphics*.

Viégas, F. B., and Wattenberg, M. (2008). *Tag clouds and the case for vernacular visualization. interactions 15*(4): 49–52.

Viégas, F. B., Wattenberg, M., and Feinberg, J. (2009). Participatory visualization with wordle. *TVCG: Transactions on Visualization and Computer Graphics*.

Vuillemot, R., Clement, T., Plaisant, C., and Kumar, A. (2009). What's being said near 'martha'? exploring name entities in literary text collections. In *VAST 2009: IEEE Symposium On Visual Analytics Science And Technology.*

Wattenberg, M., and Viégas, F. B. (2008). The word tree, an interactive visual concordance. *TVCG: Transactions on Visualization and Computer Graphics 14*(6): 1221–1228.